HANDOUT #2 - TYPES OF STATISTICAL STUDIES

TOPICS

1. Observational vs Experimental Studies
2. Retrospective vs Prospective Studies
3. Sampling Principles:
   (a) Probability Sampling: SRS, Systematic, Stratified, Cluster
   (b) Estimation of population parameters
4. Experimental Design Principles
5. Common Problems in Designed Experiments
6. Selecting an Appropriate Design
Two basic types of studies: Observational and Experimental

- Observational Study: Records information about subjects without applying any treatments to subjects (passive participation of researcher)

- Experimental Study: Records information about subjects while applying treatments to subjects and controlling study conditions to some degree (active participation of researcher)

Observational studies are of three basic types:

- Sample Survey: Provides information about a population based on a sample from the population at a specific time point.

- Prospective Study: Observes population in the present by using a sample survey and proceeds to follow the sample forward in time in order to record the occurrence of specific outcomes.

- Retrospective Study: Observes population in the present by using a sample survey and collects information about the sample about the occurrence of specific outcomes that have already taken place.

Comparison of Retrospective and Prospective Studies

- Retrospective studies are generally cheaper and can be completed more rapidly than prospective studies.

- Retrospective studies have problems due to inaccuracies in data due to recall errors.

- Retrospective studies have no control over variables which may affect disease occurrence.

- In prospective studies subjects can keep careful records of their daily activities

- In prospective studies subjects can be instructed to avoid certain activities which may bias the study

- Although prospective studies reduce some of the problems of retrospective studies they are still observational studies and hence the potential influences of confounding variables may not be completely controlled. It is possible to somewhat reduce the influence of the confounding variables by restricting the study to matched subgroups of subjects.

- Both prospective and retrospective studies are often comparative in nature. Two specific types of such studies are cohort studies and case-control studies.

  - Cohort Studies: Follow a group of subjects forward in time to observe the differences in characteristics of subjects who develop a disease with those who do not. (In place of disease, we could observe which subjects commit a crime or become PhD’s in statistics.)

  - Case-Control Studies: Identify two groups of subjects, one with the disease and one without the disease. Next, gather information about the subjects from their past concerning risk factors which are associated with the disease.
SAMPLING PRINCIPLES

Suppose a researcher wants to make an inference about a specific population. They may choose to inspect a small portion of the population, a sample. Alternatively, they could perform a census, that is, an inspection of the entire population.

Why select a sample in place of a census?

- Reduced cost
- Less time consuming
- More information per subject - Less effort expended per sampling unit
- Greater accuracy - better training of technicians, more accurate measurements, subjects may be missed in census

Sampling Frame A complete list of all \( N \) units in the population

Note: There is a 1-1 correspondence between the numbers 1, 2, \ldots, \( N \) and the sampling frame.

Probability Sampling

1. Given a frame, one can define all the possible samples that could be selected from the population. Label the distinct samples \( S_1, S_2, \ldots, S_k \).

2. Assign a probability \( S_i, P(S_i) \), to each possible sample \( S_i \), \( \sum_{i=1}^{k} P(S_i) = 1 \).

3. The sample is selected by using a random process in which the sample \( S_i \) has probability \( P(S_i) \) of being chosen.

Advantage of probability sampling: allows an objective assessment of the accuracy of inferences made about the population based on the information in the sample.

Example of non-probability sampling:

1. Convenience Sample: Data selected based on the availability of data. Historical data, medical records, production records, student academic records, select next 50 people who walk in store are examples of convenience samples.

   Problems: Data may yield a sample which is not representative of the population due to many uncontrolled variables which may be confounded with the sampling strategy.

2. Judgemental Sample: an expert selects “typical” or “representative” members of the population.

   Problem: This type of process is extremely subjective and does not admit a scientific assessment of accuracy.
SIMPLE RANDOM SAMPLING (SRS)

SRS is the most basic method of taking a probability sample. In this method of selecting a sample of \( n \) units from a population of \( N \), each of the \( \binom{N}{n} \) possible samples has the same chance of being selected. The actual choice of a specific sample can be done using a random number generator on a computer. The following S+ commands can be used.

The following commands generate random permutations of \( n \) integers or random sample from a population of numbers.

1. Random permutation of integers 1 to \( n \) : "sample(n)"
   
   EX. `sample(10)`
   
   3 8 10 6 9 5 1 4 7 2

2. Random permutation of elements in a vector \( x \) : "sample(x)"

   EX. `x<-c(23, 45, 67, 1, -45, 21, .9, 4, -3, .25)`
   `sample(x)`

   -3.00 45.00 21.00 0.90 0.25 23.00 67.00 4.00 -45.00 1.00

3. Random sample of \( n \) items from \( x \) without replacement: "sample(x,n)"

   EX. `sample(x,5)`

   67.00 21.00 45.00 0.25 -45.00

4. Random sample of \( n \) items from \( x \) with replacement: "sample(x,n,replace=T)"

   EX. `sample(x,5,replace=T)`

   -45.0 4.0 -3.0 -45.0 0.9

5. Random sample of \( n \) items from \( x \) with elements of \( x \) having differing probabilities of selection: "sample(x,n,replace=T,p)", where \( p \) is a vector of probabilities, one for each element in \( x \).

   EX. `x<-c(23, 45, 67, 1, -45, 21, .9, 4, -3, .25)`
   `p<-c(.1, .1, .1, 0, 0, 0, 0, 0, .7)`
   `sample(x,5,replace=T,p)`

   0.25 0.25 45.00 0.25 0.25

6. Randomly select \( n \) integers from the integers 1 to \( N \), without replacement:

   "sample(N,n)"

   EX. `sample(1000,10)`

   189 182 638 903 112 126 490 928 850 291
SYSTEMATIC RANDOM SAMPLING

Suppose we have a list of the population units or units are produced in a sequential manner. A 1-in-$k$ systematic sample consists of selecting one unit at random from the first $k$ units and then selecting every $k$th unit until $n$ units have been collected. In a population containing $N$ units, systematic sampling has a selection probability of $\frac{n}{N}$ for each unit. However, not all $\binom{N}{n}$ possible samples are equally likely, as in SRS.

Systematic sampling is often used when a sequential list of sampling units exists or when sampling units become available in a sequential manner. Systematic sampling provides a sample which is representative of the population provided there are no cyclic patterns in the population lists.

STRATIFIED RANDOM SAMPLING

Population is divided into $L$ groups or strata. The strata are nonoverlapping and contain $N_1, N_2, \ldots, N_L$ units respectively. Note: $N_1 + N_2 + \cdots + N_L = N$. Suppose simple random samples of sizes $n_1, n_2, \ldots, n_L$ are selected independently from the $L$ strata. This sampling procedure is known as stratified random sampling.

Reasons for Using a Stratified Random Sample:

- Precise estimates within subpopulations (strata)
- Administrative convenience
- Sampling problems differ according to different parts of the population.
- Possible gain in precision in the overall estimate of population parameters. (This occurs when there are large differences between stratum but there is homogeneity within the $L$ strata.

SINGLE-STAGE CLUSTER RANDOM SAMPLING

Population consists of $N$ units or clusters. The $N$ clusters contain $M_1, M_2, \ldots, M_N$ smaller units called elements. Population contains a total of

$$\sum_{i=1}^{N} M_i = M^*$$

elements

Cluster Sample A SRS of $n$ cluster is selected and all elements within each cluster is measured or surveyed.

Suppose $M_i = M$ for all $i$. What advantage is there to taking a cluster sample of $nM$ elements as opposed to a SRS of $nM$ elements from the population? In general, the cluster sample will be less precise than the SRS. The main reason for using cluster sampling is administrative difficulties of obtaining a frame for all $M^*$ elements in the population. For example, suppose an element is a household in Houston. Define a cluster as a city block in Houston. Obtaining a frame of all city blocks in Houston is undoubtedly easier than obtaining a frame of all households in Houston.
ESTIMATION OF POPULATION MEAN: $\mu$

Consider the estimation of $\mu$ under three different sampling Methods

**Simple Random Sampling**

Let $y_1, y_2, \ldots, y_n$ be the measurements obtained from the SRS of $n$ units from the population. The estimator of population mean $\mu$ is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

with estimated variance of $\hat{\mu}$ given by

$$\text{Var}(\hat{\mu}) = \frac{s^2}{n} \left( \frac{N - n}{N} \right)$$

where $s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2$.

**Stratified Random Sampling**

Suppose we have independently selected SRS’s of size $n_1, n_2, \ldots, n_L$ from the $L$ strata. Let $\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_L$ be the sample means of the $L$ SRS samples selected from the $L$ strata with the number of units in each strata given by $N_1, N_2, \ldots, N_L$. The estimator of population mean $\mu$ is

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{L} N_i \bar{y}_i$$

with estimated variance of $\hat{\mu}$ given by

$$\text{Var}(\hat{\mu}) = \frac{1}{N^2} \left[ \sum_{i=1}^{L} N_i^2 \left( \frac{N_i - n_i}{N_i} \right) s_i^2 \right]$$

where $s_i^2 = \frac{1}{n_i-1} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2$.

**SINGLE STAGE CLUSTER Random Sampling**

Let $N$ be the number of clusters in the population; $n$ be the number of clusters selected in a simple random sample from the population; $m_i$ be the number of elements in cluster $i$, $i = 1, 2, \ldots, N$; $\bar{m} = \frac{1}{n} \sum_{i=1}^{n} m_i$ be the average cluster size for the sample of $n$ clusters, $M = \sum_{i=1}^{N} m_i$ be the number of elements in the population, $\bar{M} = \frac{M}{N}$ be the average cluster size for the population, $y_i = \sum_{j=1}^{m_i} y_{ij}$ be the total of all measurements of the $m_i$ elements in the $i$th cluster. The estimator of population mean $\mu$ is

$$\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$$

with estimated variance of $\hat{\mu}$ given by

$$\text{Var}(\hat{\mu}) = \left( \frac{N - n}{N n M^2} \right) \sum_{i=1}^{n} (y_i - m_i \hat{\mu})^2 \frac{1}{n - 1}$$
EXPERIMENTAL DESIGN PRINCIPLES

Selected Comments from *Experimental Design* by W. Federer

I. “All fields of research have at least one feature in common:
   The variability of experimental responses.”

II. When there is considerable variation from observation to observation on the same experimental material and it is not feasible to run a large number of experiments (which would reduce the variation in the mean response), THEN the experimenter must:

   1. Refine the experimental design in order to obtain a specified degree of precision (Blocking)
   2. In order to attach a probability statement to the observed treatment mean differences (a measure of the degree of confidence in the observed results), it is necessary that proper Randomization and Replication occur.

III. Certain Principles of Scientific Experimentation should alway be followed: (Many are nonstatistical, however, the analysis of the data resulting from improperly designed and conducted experiments may complicate the analysis to the point at which NO analysis of the data can be conducted.)

   P1. Formulation of Questions to be Asked and Research Hypotheses to be Tested:

      Clearly stating and precisely formulating questions and hypotheses prior to the running of the experiments will help to
      1. Minimize the number of replications required
      2. Make sure all necessary measurements are taken.

   P2. A Critical and Logical Analysis of the Stated Research Hypotheses:

      1. Review the relevant literature
      2. Evaluate the reasonableness and utility of the aim of the experiment as reflected in the Research Hypotheses. (May need to reformulate the Research Hypotheses.)
      3. Forecast the possible outcomes of the experiment in order to determine if the resulting data can be analyzed using the proper statistical methodology: For example,
         Too many 0’s,
         Categorical data,
         Too few replications for projected variability,
         Correlated (nonindependent observations)

   P3. Selection of Procedures for Conducting Research

      1. What Treatments to be included in experiment?
      2. What Measurements should be made on the experimental units?
      3. How should experimental units be selected?
      4. How many experimental units should be used?
      5. What sampling or experimental design should be used?
      6. What is the effect of adjacent experimental units on each other? How can this effect be controlled? (Competition between experimental units leads to dependent data.)
      7. Outline of pertinent summary tables for recording data.
      8. Experimental procedures outlined and documented.
      9. Statement of costs in terms of materials, personnel, equipment.
10. Consideration of the above items may often result in a restructured experiment, rather than an experiment in which the results are highly incomplete and not very useful.

P4. Selection of suitable Measuring Devices and Elimination of Personal Biases and Favoritisms:
1. Never observe 3 samples and discard “most discrepant” observation
2. Never place “Favorite Treatment” under the best experimental conditions
3. Discard 0’s or values from abnormal experimental units only after a critical examination of the experimental units and a determination of the degree of unsuitability of the results in reference to standard experimental conditions. Always report the data values and explain why they were excluded from the analysis.

P5. Carefully evaluate the statistical tests and the necessary conditions needed to apply these tests with respect to experimental procedures and underlying distributional requirements. (Residual analysis to check that assumptions hold.)

P6. Quality of the Final Report:
1. Include well designed graphics
2. Include description of statistical procedures and data collection methodology so that the reader of the report can determine the validity of your experiment and analysis.
3. Report should be prepared whether or not the research hypotheses have been supported by the data; otherwise Type I errors alone may produce misleading conclusions. Many experiments result in the acceptance of the null hypothesis but no report is written. Thus, even when the research hypothesis is in fact false but many experiments were conducted concerning this hypothesis, there may be a number of these experiments (5% Type I Errors) that support this research hypothesis incorrectly whereas a large number of experiments (95%) in fact find that the research hypothesis is not supported by the data but since report is written the research hypothesis may be incorrectly supported in the literature.
4. It is crucial that the size of the treatment effect, for example an estimate of $\mu_i - \mu_i'$, be reported and not just the p-value of the test. Include confidence intervals on the effect size. Thus, a distinction is being made between Statistically Significant Results (small p-value) and Practically Significant Results (small p-value with large Treatment effect).

IV. Statistically Designed Experiments are
- Economical
- Allow the measurement of the influence of several factors on a response
- Allow the estimation of the magnitude of experimental variability
- Allow the proper application of statistical inference procedures
EXPERIMENTAL DESIGN TERMINOLOGY

I. Designed Experiment Consists of Three Components:

C1. Method of Randomization:
   a. Completely Randomized Design (CRD)
   b. Randomized Complete Block Design (RCBD)
   c. Balanced Incomplete Block Design (BIBD)
   d. Latin Square Design
   e. Crossover Design
   f. Split Plot Design
   g. Many others

C2. Treatment Structure
   a. One Way Classification
   b. Factorial
   c. Fractional Factorial
   d. Fixed, Random, Mixed factor levels

C3. Measurement Structure
   a. Single measurement on experimental unit
   b. Repeated measurements on experimental unit: Different Treatments
   c. Repeated measurements on experimental unit: Longitudinal or Spatial
   d. Subsampling of experimental unit

II. Specific Terms Used to Describe Designed Experiment:

1. Experimental Unit: Entity to which treatments are randomly assigned
2. Measurement Unit: Entity on which measurement or observation is made (often the experimental units and measurement units are identical)
3. Homogeneous Experimental Unit: Units that are as uniform as possible on all characteristics that could affect the response
4. Block: Group of homogeneous experimental units
5. Factor: A controllable experimental variable that is thought to influence the response
6. Level: Specific value of a factor
7. Experimental Region (Factor Space): All possible factor-level combinations for which experimentation is possible
8. Treatment: A specific combination of factor levels
9. Replication: Observations on two or more units which have been randomly assigned to the same treatment
10. Subsampling: Multiple measurements (either longitudinally or spatially) on the same experimental unit under the same treatment
11. Response: Outcome or result of an experiment
12. Effect: Change in the average response between two factor-level combination or between two experimental conditions
13. Interaction: Existence of joint factor effects in which the effect of each factor depends on the levels of the other factors
14. Confounding: One or more effects that cannot unambiguously be attributed to a single factor or interaction
15. Covariate: An uncontrollable variable that influences the response but is unaffected by any other experimental factors
A semi-conductor manufacturer is having problems with scratching on their silicon wafers. They propose applying a protective coating to the wafers, however, the wafer engineers are concerned about the diminished performance of the wafer. An experiment is designed to evaluate several types and thicknesses of coatings on the conductivity of the wafer. Two types of coatings and three thicknesses of the coating are selected for experimentation. A random sample of 72 wafers are selected for use in the experiment with 12 wafers randomly assigned to each combination of a type of coating ($C_1, C_2$) and a thickness of coating ($T_1, T_2, T_3$). Only 24 wafers can be evaluated on a given day, thus it will take 3 days to complete the experiment. On each wafer, the conductivity is recorded before and after applying the coating to the wafer. Furthermore, to assess the variability in conductivity across the wafer surface, conductivity readings are taken at five locations on each wafer.

For the above experiment, identify all the components and definitions:

I. Designed Experiment Consists of Three Components:

C1. Method of Randomization:

C2. Treatment Structure:

C3. Measurement Structure:

II. Specific Terms Used to Describe Designed Experiment:

1. Experimental Unit:

2. Measurement Unit:

3. Homogeneous Experimental Unit:

4. Blocks:

5. Factors:
6. Levels:

7. Experimental Region (Factor Space):

8. Treatment:

9. Replication:

10. Subsampling:

11. Response:

12. Effect:

13. Interaction:

14. Confounding:

15. Covariate:
COMMON PROBLEMS IN EXPERIMENTAL DESIGNS

I. Masking of Factor Effects

When the variation in the responses are as large as the differences in the treatment means, the treatment differences will not be detected in the experiment. For example, \( \sigma_e \) is large relative to \( \mu_i - \mu_{i'} \) in a completely randomized design. In this situation, the experiment must be redesigned by

1. Increasing the sample sizes to reduce
   \[ \text{StDev}(\hat{\mu}_i - \hat{\mu}_{i'}) = \sigma_e \sqrt{\frac{1}{n_i} + \frac{1}{n_{i'}}} \]
2. Blocking the experimental units to reduce the size of \( \sigma_e \)
3. Using Covariates
4. All the above

II. Uncontrolled Factors
If factors are known to have an effect on the response variable, then these factors should be included in the experiment as either treatment or blocking variables. Failure to carefully consider all factors of importance can greatly compromise the extent to which conclusions can be drawn from the experimental outcomes.

1. Differences between experimental plots in terms of soil fertility, drainage, exposure to sun, exclusion of wildlife, etc.
2. Position of experimental units on greenhouse benches
3. Position of experimental units on trays or in ovens
4. Time of day or week in which experiment is run

III. Erroneous Principles of Efficiency
If the time to run experiments or the cost to run experiments place restrictions on the number of factors and the number of levels of the factors that can be included in the experiment, then the overall goals of the experiment must be reevaluated since

1. Important factors may be ignored or left uncontrolled
2. Non-linear effects may not be determined since the number of levels may be too few or not broad enough to detect higher order effects.

Conductivity Related to Thickness of Silicon Wafers Coating

<table>
<thead>
<tr>
<th>Thickness of Coating (mm)</th>
<th>Conductivity of Wafer (ohms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
</tr>
<tr>
<td>6</td>
<td>35</td>
</tr>
<tr>
<td>7</td>
<td>40</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
</tr>
<tr>
<td>9</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>55</td>
</tr>
</tbody>
</table>

Conductivity Related to Thickness of Silicon Wafers Coating

Thickness of Coating (mm)
SELECTING AN APPROPRIATE EXPERIMENTAL DESIGN

I. Consideration of Objectives

1. Nature of anticipated results helps to determine what factors need to be included in the experiment: Suppose experiment is designed to determine which of 6 fuel blends used in automobiles produce the lowest CO emissions. The 6 blends include a standard commercial gasoline and 5 different methanol blends. After determining that blend number 5 has the lowest CO emission, the question arises what properties of the blends (distillation temperature, specific gravity, oxygen content, etc.) made the major contributions to the reduced CO level in emissions using the selected blend. A problem that may arise is that the fuel properties may be confounded across the 5 blends and it may not be possible to sort them out with the given experimental runs. This problem could have been avoided if this question was raised prior to running the experiments.

2. Definition of concepts (Can the goals of the experiment be achieved) : Suppose we want to study the effects of radiation exposure on the life length of humans
   - Design 1: Subject randomly selected homogeneous groups of humans to various levels of radiation (unethical experiment)
   - Design 2: Use laboratory rats in place of humans (extrapolation problem)
   - Design 3: Use observational or historical data on groups that were exposed to radiation (Many uncontrolled factors, genetic differences, amount of exposure, length of exposure, occupational differences, daily habits)

3. Determination of observable variables What covariates should be observed? How often? How accurately should they be measured?

II. Factor Effects

1. Inclusion of all relevant factors avoids uncontrolled systematic variation.
2. Need to measure all important covariates to control heterogeneity of experimental units or conditions.
3. Anticipated interrelationships between factor levels helps to determine type of design:
   a. No interactions between factor levels: Use simple screening design
   b. Interactions exist: Need full factorial design
   c. Higher order relationships between factor levels may require a greater number of levels of the factors in order to be able to fit high order polynomials to the responses.
4. Include a broad enough range of the factor levels so as not to miss important factor effects, include lowest and highest feasible values of factor.

III. Precision - Efficiency of Experiment Degree of variability in response variable determines the number of replications required to obtain desired widths of confidence intervals and power of statistical tests. Determine variability through pilot studies or review literature for results from similar experiments.

IV. Randomization

In order to protect against unknown sources of biases and to be able to conduct valid statistical procedures:

1. The experimental units MUST be randomly assigned to the treatments or
2. The experimental units MUST be randomly selected from the treatment populations and
3. The time order in which experiments are run and/or spatial positioning of experimental units must be randomly assigned to the various treatments. This avoids the confounding of uncontrolled factor effects with the experimental factors. For example, drifts in instrumental readings, variation across the day in terms of temperature gradients, humidity or sunlight exposure, variation in performance of laboratory technicians (grad students), or various other conditions in the laboratory or field.

**DESIGNING FOR QUALITY: INDUSTRIAL PROCESSES**

Two Basic Types of Experiments

1. **On-Line**: Running experiments while process is in full production. EVOP - Evolutionary Operation Design strategy where 2 or more factors in an on-going production process are varied in order to determine an optimal operation level. Problem: Examining very narrow region of the factor space since only small deviations from normal operations are allowed by the company.

2. **Off-Line**: Running experiments in Laboratories or Pilot Plants

Two Basic Goals in Experiments Involving Quality Improvement

1. **Bring product On Target** Average measurement of product characteristic are equal to the target value

2. **Uniformity - Consistency** Measured product characteristics have a small variability about the target value

Combining both of these criterions, we obtain

\[
\text{Minimize } \text{MSE} = (\text{Bias})^2 + (\text{StDev})^2 = (\text{Distance to Target})^2 + \text{Variance}
\]

**Taguchi Approach:**

1. Emphasized the importance of using fractional factorial designs

2. His choice of designs were often highly inefficient

3. His analyses of experiments were often incorrect

4. He was successful in convincing engineers at large corporations to use designed experiments. The experiments were very successful even though there were not the best possible experiments that could have been run.